Analysis of Expressed Sequence Tags From Two Starvation, Time-of-Day-Specific Libraries of *Neurospora crassa* Reveals Novel Clock-Controlled Genes

Hua Zhu,*[,1] Minou Nowrousian,[†,1] Doris Kupfer,* Hildur V. Colot,[†] Gloria Berrocal-Tito,[†] Hongshing Lai,* Deborah Bell-Pedersen,[‡] Bruce A. Roe,* Jennifer J. Loros[†] and Jay C. Dunlap[†]

**Department of Chemistry and Biochemistry, Advanced Center for Genome Technology, University of Oklahoma, Norman, Oklahoma 73019, [†]Departments of Genetics and Biochemistry, Dartmouth Medical School, Hanover, New Hampshire 03755 and [‡]Department of Biology, Biological Sciences, Texas A&M University, College Station, Texas 77843*

Manuscript received October 11, 2000
Accepted for publication December 19, 2000

ABSTRACT

In an effort to determine genes that are expressed in mycelial cultures of *Neurospora crassa* over the course of the circadian day, we have sequenced 13,000 cDNA clones from two time-of-day-specific libraries (morning and evening library) generating ~20,000 sequences. Contig analysis allowed the identification of 445 unique expressed sequence tags (ESTs) and 986 ESTs present in multiple cDNA clones. For ~50% of the sequences (710 of 1431), significant matches to sequences in the National Center for Biotechnology Information database (of known or unknown function) were detected. About 50% of the ESTs (721 of 1431) showed no similarity to previously identified genes. We hybridized Northern blots with probes derived from 26 clones chosen from contigs identified by multiple cDNA clones and EST sequences. Using these sequences, the representation of genes among the morning and evening sequences, respectively, in most cases does not reflect their expression patterns over the course of the day. Nevertheless, we were able to identify four new clock-controlled genes. On the basis of these data we predict that a significant proportion of the expressed Neurospora genes may be regulated by the circadian clock. The mRNA levels of all four genes peak in the subjective morning as is the case with previously identified *ccgs*.

THE ascomycete *Neurospora crassa* has a long history as a model organism for both classical and molecular genetics in general (Davis 2000) as well as for investigation of circadian rhythmicity (Dunlap 1999). Several efforts have been made recently to map and sequence the Neurospora genome (Fungal Genome Resources at http://gene.genetics.uga.edu/ and Munich Information Centre for Protein Sequences *Neurospora crassa* Database at http://www.mips.biochem.mpg.de/proj/neurospora/) as well as to address the expression of genes at different stages of the Neurospora developmental cycle (Neurospora Genome Project at http://www.unm.edu/~ngp/; Nelson *et al.* 1997; Dolan *et al.* 2000).

One of the most notable aspects of *N. crassa* is the fact that a great part of its life is tightly controlled by the circadian clock. Not only the formation of macroconidia but also the expression of genes important for general metabolism occur in a circadian fashion (Loros 1998). Using differential screening procedures of cDNA libraries, several genes have been shown to be under clock control (Loros *et al.* 1989; Bell-Pedersen *et al.* 1996). These studies have confirmed the impression that a substantial proportion of Neurospora genes might be regulated by the circadian clock, but the extent of this control is yet unclear. Therefore, it was decided to use an expressed sequence tag (EST) sequencing approach to address the question of clock regulation at a larger scale as well as to identify novel ESTs and to examine global gene expression under conditions (starvation in the dark) not previously examined.

EST sequences are highly valuable in genomic approaches. EST data can be used to identify novel genes, search for homologous genes in different organisms, analyze alternative splicing, and identify chromosomal locations of genes and other applications (Pandey and Lewitter 1999; Ohlrogge and Benning 2000). With large numbers of ESTs available, so-called "digital Northerns" or "electronic transcriptional profiling" can be performed by counting the number of ESTs for a given gene within the EST population (Pandey and Lewitter 1999; Prade *et al.* 2001). Also, cDNA clones with known sequences are used to establish cDNA microarrays with which genome-wide expression patterns can be investigated (Duggan *et al.* 1999).

In the project described here, cDNA clones from two mycelial libraries of *N. crassa* were partially sequenced. The libraries were made from time-of-day-specific tissues and had been used to characterize clock-controlled genes (*ccgs*) earlier (Bell-Pedersen *et al.* 1996). The ESTs were organized into contiguous sequences (contigs)

Corresponding author: Jay C. Dunlap, Department of Genetics, Dartmouth Medical School, Hanover, NH 03755.
E-mail: jay.c.dunlap@dartmouth.edu

[1]These authors contributed equally to this work.

and compared to the sequences within the National Center for Biotechnology Information (NCBI) database. Besides genes that have homologs in the NCBI database, the sequencing yielded a large number of yet unknown ESTs (50% of 1431 different genes identified within the libraries). Northern blots were hybridized with probes derived from 26 cDNA clones, and 4 of these genes were found to be rhythmically expressed. Again, this confirms the prediction that a significant proportion of *N. crassa* genes may be regulated by the circadian clock.

FIGURE 1.—Derivation of ESTs. Sequences generated using the forward and reverse primers correspond to the 5′ and 3′ ends of the cDNA clones and are denoted with f1 and r1, respectively, at the end of their sequence identification number. cDNAs were directionally cloned using asymmetric primers and linkers as described in BELL-PEDERSEN *et al.* (1996).

MATERIALS AND METHODS

Strains and growth conditions: The following *N. crassa* strains were used: *frq*$^{+}$ strain 87-3 (*bd; a*) and long period mutant 585-7 (*bd; a; frq*^{7}). Neurospora media (Vogel's) were as described (DAVIS and DESERRES 1970). Culture conditions for rhythmic RNA analysis and light induction experiments were performed according to published methods (LOROS *et al.* 1989).

Preparation and analysis of RNA: RNA was prepared as described previously (YARDEN *et al.* 1992). Northern blots, slot blots, and probing with DNA probes or riboprobes were performed according to standard techniques (MANIATIS *et al.* 1982).

Template preparation for DNA sequencing: The libraries from which the sequenced clones are derived have been described previously (BELL-PEDERSEN *et al.* 1996); they were not further amplified or modified except for those modifications that normally arise during the process of phagemid excision (see below). Template preparation, sequencing, and analysis were performed at the Advanced Center for Genome Technology, University of Oklahoma. For each library, infection of *Escherichia coli* strain SOLR with aliquots from the mass library excision converted the single-stranded phagemids to the double-stranded form. The appropriate volume of F_1 lysate containing the single-stranded phagemids was added to the SOLR cells to yield 200–300 colonies per 2×10^7 of infected SOLR cells. Single white colonies were picked to individual wells of a 96-well block containing 1.5 ml TB with ampicillin per well. After growth, triplicate 100-μl aliquots of each cell culture were mixed with glycerol (final concentration 17%) and stored at −70°. Individual clones are available publicly from the Fungal Genetics Stock Center, University of Kansas Medical Center (Kansas City, KS; http://www.fgsc.net). Double-stranded DNA template was robotically isolated from the remaining cell culture using a modified plasmid preparation procedure (http://genome. ou.edu/proto.html; BIRNBOIM and DOLY 1979) on a Biomek 2000 (Beckman, Fullerton, CA).

DNA sequencing: Partial nucleotide sequences of the cDNA inserts were determined by single-pass sequencing using the Big Dye sequencing system (Perkin Elmer, Norwalk, CT; ROSENBLUM *et al.* 1997) with either the universal forward (5′ GACGTTGTAAAACGACGGCC) or the universal reverse primer (5′ CACAGGAAACAGCTATGACC). Generally, 0.2 μg of each template was cycle sequenced with 6.5 pmol primer and 2 μl of a 1:4 diluted AmpliTaq DNA polymerase Big Dye termination mix in a final volume of 5–7 μl. Cycling conditions were 60 cycles of 95° for 10 sec, 50° for 5 sec, and 60° for 4 min in Perkin Elmer thermocyclers 9600 or 9700. Reactions were purified by centrifugation through Sephadex G-50 microtiter plate spin columns (BODENTEICH *et al.* 1994) and electrophoresed on an ABI 377 sequencer (Applied Biosystems, Foster City, CA) for 7 hr at 2.4 kV. The collected raw data was manually retracked and analyzed using the ABI sequencing analysis software. See http://www.genome.ou.edu/proto.html for more detailed descriptions of protocols.

Informatics: DNA sequences were automatically surveyed for quality using the phred software (EWING *et al.* 1998). Sequences were screened for the presence of vector sequences, low quality sequences (Phred-16), and small inserts (<100 bp) and were trimmed to remove poly(A) tail and vector sequences. All sequences were screened for homology to ribosomal RNA, mitochondrial DNA, and *E. coli* genomic sequences using BLASTN (ALTSCHUL *et al.* 1997) and matches were removed. To reduce the redundancy of the libraries and determine the relative abundance of the members, overlapping cDNA sequences were identified and organized into contigs using the phrap software (P. Green, copyright 1994–1996, http://bozeman.mbt.washington.edu/phrap.docs/phrap.html) with a window size of 14 and a minimal score of 80. A contig here, then, arises from the alignment of clustered ESTs on the basis of sequence similarity. The derived peptide sequences corresponding to the cDNA sequences were compared with the protein database available through the NCBI (Bethesda, MD) using the BLASTX algorithm (ALTSCHUL *et al.* 1997). EST sequences and results of comparisons can be obtained through the website http://www.genome.ou.edu/fungal.html. The site also provides the opportunity to search the EST sequences using BLAST. The EST sequences were also deposited in the dbEST database of GenBank (BOGUSKI *et al.* 1993).

RESULTS

Sequence analyses: Plasmid DNA from >13,000 clones derived from two *N. crassa* cDNA libraries was sequenced using the universal reverse or universal forward primer (Figure 1). The libraries have been used previously to identify clock-controlled genes (*ccgs*) and correspond to "morning" (circadian time CT1) and "evening" (CT13; BELL-PEDERSEN *et al.* 1996). For >53% of all clones, high-quality sequence data (see MATERIALS AND METHODS) were obtained with both primers, 28% of all clones were sequenced only with the forward primer, and 19% only with the reverse primer. Typically 300–500 bp of high-quality sequence was obtained. Sequences obtained with the reverse primer should correspond to the 3′ end and are designated in the database with an "r1" at the end of the sequence identification number; sequences obtained with the forward primer should correspond to the 5′ end and have "f1" at the end of their identification number. Exceptions occurred for those few clones in which the cDNA inserts were inserted

in reverse orientation. The sequences are available in a publicly searchable format as described in MATERIALS AND METHODS.

Contig analysis and cDNA identities: We obtained 9148 sequences from the evening and 10,871 sequences from the morning library, representing a total of 1431 different genes. Of these genes, 445 are represented only once within the libraries, whereas 986 are represented by more than one EST. The EST sequences were organized into contigs using the phrap software (P. Green, copyright 1994–1996, http://bozeman.mbt.washington.edu/phrap.docs/phrap.html). To identify potential homologs to the Neurospora genes, the BLASTX algorithm (ALTSCHUL *et al.* 1997) was used to translate each contig nucleotide sequence into the six possible reading frames and to compare the predicted protein sequences with the NCBI protein sequence database. In previously published sequence analyses of Neurospora ESTs (NELSON *et al.* 1997), matches (putative homologs) were divided into highly significant (P/E values $\leq 10^{-20}$), moderately significant (10^{-5}–10^{-19}), and weakly significant (10^{-2}–10^{-4}). Here, we present only those matches that are at least moderately significant. The data upon which this analysis is based can be viewed at the Genetics website at http://www.genetics.org/supplemental/, http://www.genome.ou.edu.fungal.html, and http://www.dartmouth.edu/~jdunlap/. The clones are shown according to the classification scheme utilized by the Expressed Gene Anatomy Database (EGAD, available at http://www.tigr.org/docs/tigr-scripts/egad_scripts/role_report.spl; WHITE and KERLAVAGE 1996). A total of 710 of 1431 sequences (50%) correspond to characterized genes, including genes of known function as well as putative open reading frames. The remaining 721 genes (50%) do not have any significant matches within the NCBI database. Of the total set of 1431 genes, ~35% were represented in the New Mexico Neurospora EST database (NELSON *et al.* 1997), and of these about half had sequence homologs in the NCBI database and half did not.

Transcript abundance in the morning and evening libraries: With a large number of ESTs available, so-called "digital Northerns" or "electronic transcriptional profiling" can be performed by counting the number of ESTs for a given gene within an EST population. Formulas have been developed to estimate the statistical significance of this electronic transcript profiling (AUDIC and CLAVERIE 1997), but few efforts have been made to evaluate the correlation between digital Northerns and laboratory methods for the analysis of gene expression. Several reports describe approaches to correlate these data for human or plant EST databases (HUANG *et al.* 1999; BORTOLUZZI *et al.* 2000; HWANG *et al.* 2000; MEKHEDOV *et al.* 2000), but no data have been published previously to achieve this for fungal gene expression profiling (see, however, PRADE *et al.* 2001). Therefore, efforts were made to evaluate whether the relative abundance of cDNAs in the morning *vs.* the evening library reflects the situation *in vivo*. Insert DNAs of 26 cDNA clones representing contigs that contain multiple cDNA clones were used to hybridize Northern or slot blots with RNA from time courses. Growth and harvest of tissues was performed as described previously (BELL-PEDERSEN *et al.* 1996); the clones used as probes are given in Table 1. Among the 26 genes represented by the cDNAs, 19 should be differentially expressed with $P > 0.999$ according to published formulas (http://igs-server.cnrs-mre.fr; AUDIC and CLAVERIE 1997). Two genes have probabilities of being differentially expressed between 0.95 and 0.999, and 5 have probabilities of <0.95; cDNAs representing these latter ones were chosen as "negative controls" (Table 1). These results indicate that preferential occurrence in either the morning or evening library did not generally correspond to rhythmic expression *in vivo* or to enhanced expression at the time of day indicated by the occurrence of transcripts in the libraries (Table 1). The majority of the genes investigated showed more or less even expression over the course of the circadian day or displayed small peaks at random times (data not shown). Also, previously characterized *ccgs* (BELL-PEDERSEN *et al.* 1996) were in most but not all cases represented in the libraries according to their expression patterns *in vivo* (Table 2). As has been shown previously, *ccgs* peak preferentially in the late night or early morning (BELL-PEDERSEN *et al.* 1996), but this pattern could be found within the EST sequences only in five out of seven *ccgs* (Table 2). Possible explanations for these findings are discussed below (see DISCUSSION).

Identification of clock-controlled genes: Of the 26 cDNA clones used to probe Northern or slot blots, 4 displayed rhythmic expression patterns (Table 3, Figure 2). These were named *ccg-13*, *ccg-14*, *ccg-15*, and *lyz* and corresponded to contigs 1432, 1421, 1411, and 1442 (Table 3). Probes derived from representative cDNA clones were hybridized to RNA from both frq^+ and frq^7 cultures to verify that the period length of the rhythm is dependent on the *frq* allele (21.5 hr for frq^+ and 29 hr for frq^7) and, therefore, that the expression is clearly under clock control. As shown in Figure 2, expression of *ccg-13* and *ccg-14* was cyclic, displaying high amplitudes, and the troughs of their rhythms were very consistent (Figure 2B). The other two genes, *ccg-15* and *lyz*, displayed much lower amplitudes, and the variation between different assays appears to be much greater (Figure 2B). Nevertheless, a consistent pattern of peaks and troughs could be observed in both frq^+ and frq^7 strains for *ccg-15* as well as *lyz*, and although clock control is not as strong as with other *ccgs*, the expression of these genes appears to be clock regulated. Expression of the new *ccgs* peaked between CT0 and CT6; therefore, they were morning specific as are previously identified *ccgs* (BELL-PEDERSEN *et al.* 1996). The newly identified *ccgs* were tested for light induction by hybridization to RNAs

TABLE 1

Clones used for hybridizing time-specific RNAs

Contig no.[a]	Clone identity	Homology	Morning cDNAs	Evening cDNAs	Rhythmic expression
1314	a8d2ne	eIFA5	2	42	No
1339	a6h7ne[b]		0	32	No
1378	c8a3ne	Histone H4	50	8	No
1382	c7h2ne	PEPCK	1	48	No
1387	c8f10ne	ADH I	1	74	No
1404	b7h7ne	Rib. prot. L6	49	13	No
1405	a9d8nm	Rib. prot./Ubiquitin	51	11	No
1411	a8d1ne	SPS2 homolog	140	30	Yes (*ccg-15*)
1416	a1g7nm	Rib. prot. L14	86	4	No
1419	a2f1nm[b]		222	38	No
1422	a8e5nm	HSP70	108	6	No
1423	b8d8ne	PGK	29	93	No
1427	a8a11nm		106	3	No
1428	a1f11nm	Ubiquitin	238	42	No
1432	a5f7nm	Phase spec. protein	117	6	Yes (*ccg-13*)
1434	d3f10nm	DnaJ	134	0	No
1442	a2f10nm	Lysozyme	202	38	Yes (*lyz*)
1445	a3d4nm		108	181	No
1429c	a8g9nm	Thioredoxin	39	0	No
1334	*a8d7nm*	*RCO-3*	*76*	*38*	*No*
1415	*a2a12nm*[c]		*56*	*30*	*No*
429	b8g8ne[b]		0	3	No
814	a7h7ne[b]		1	5	No
1401	a1d7nm		37	23	No
1412	a2h10nm	V-ATPase	38	41	No
1421	a3e1nm	Snodprot1	66	41	Yes (*ccg-14*)

The identity of the clones from which the probes were derived is given in the second column, and the numbers of the contigs to which the clones are assigned are given in the first column. More detailed information about homologies is given at http://www.genetics.org/supplemental/, http://www.genome.ou.edu/fungal.html, and http://www.dartmouth.edu/~jdunlap/. Blocks of contigs are italicized or underlined according to their probabilities for differential expression as calculated from the number of ESTs present in the morning and evening libraries (Audic and Claverie 1997). The first 19 genes represented by the contigs shown below have a probability of >0.999 of being differentially expressed, the next two contigs between 0.95 and 0.999 (italicized), and the last five contigs <0.95 (underlined). PEPCK, phosphoenolpyruvate carboxykinase; PGK, phosphoglycerate kinase; V-ATPase, vacuolar ATPase.

[a] A single contig identification number is given in those cases in which several contigs have homology to the same gene. This can happen in cases when one contig corresponds to the 5′ end and one to the 3′ end of a given gene. A complete list of contigs corresponding to a single gene can be found at the Genetics website at http://www.genetics.org/supplemental/, at http://www.genome.ou.edu/fungal.html, and at http://www.dartmouth.edu/~jdunlap/.

[b] In these cases, the PCR fragments used for hybridization were not derived from the clones given, but from other clones from the same contig. The clone numbers given here are clones with significant overlap to the probes used.

[c] The clone with the sequence of a2a12nm is in Well a2b12nm of the library.

that were extracted from mycelia exposed to continuous light for intervals of 15 min to 2 hr, but none of them was light induced under these conditions (data not shown). Riboprobes were generated for both strands of the new *ccgs* and, as expected, all of them were found to encode mRNAs within their f1 strands (Figure 1).

Sequence similarities and chromosomal locations of the new clock-controlled genes: Comparison of the newly identified *ccgs* to the databases revealed the following (Table 3): *ccg-13* shows strong sequence similarity to a phase-specific protein from the ascomycete *Ajellomyces dermatitidis* (GenBank accession no. AF277086, *P/E* value $3e^{-10}$). The putative 5′ nontranslated region of contig 1432 (representing *ccg-13*) shows identity to the end of cosmid H37F12, which was sequenced by the University of Georgia mapping project (available at http://gene.genetics.uga.edu/). This cosmid was mapped to linkage group I by the same group (linkage data at http://www.fgsc.net/mappingdata/mapping1.htm); therefore, *ccg-13* should be located on linkage group I. *ccg-14* is very similar to snodprot1 from the ascomycete *Phaeosphaeria nodorum* (GenBank accession no.

TABLE 2

Abundance of previously identified *ccgs* among the EST sequences

Gene	Morning cDNAs	Evening cDNAs	Ratio morning/evening in libraries
ccg-1	351	191	1.8
ccg-2	457	248	1.8
ccg-4	0	42	—
ccg-6	33	179	0.2
ccg-7	692	214	3.2
ccg-8	37	0	—
ccg-9	66	26	2.5

The number of sequences corresponding to a certain *ccg* obtained from the morning or evening libraries is given in the second and third column, respectively, and their ratios are given in the fourth column. The ESTs for the evening library were normalized by multiplying by 1.2, as there were fewer EST sequences obtained from the evening *vs.* the morning library (9148 *vs.* 10,871).

AF074941, *P/E* value $7e^{-39}$), a protein that displays strong similarity to cerato-platanin, a phytotoxin from the ascomycete *Ceratocystis fimbriata* with N-terminal homology to the hydrophobin family (Pazzagli *et al.* 1999). It is contained in contig 9A11 from the MIPS *N. crassa* database (http://www.mips.biochem.mpg.de/proj/neurospora/), which maps to linkage group V. *ccg-15* displays strong sequence similarity to an SPS2 homolog from *Saccharomyces cerevisiae* (Saccharomyces Genome Database via RefSeq, accession no. NP_009634, *P/E* value $9e^{-45}$) and is contained in clone 3e7 from the MIPS *N. crassa* database, which also maps to linkage group V. Thus, although there are clear sequence homologs to *ccg-13*, *ccg-14*, and *ccg-15*, there is at present nothing about the biology of Neurospora to suggest biologically (or phenotypically) meaningful names for these genes, and so they have provisionally been identified as *ccg-13*, *14*, and *15*, pending functional studies. Contig 1442 shows the strongest homology to lysozyme from the ascomycete *Chalaropsis* sp. (Swissprot accession no. P00721, *P/E* value $4e^{-85}$; Felch *et al.* 1975); therefore, the gene was named *lyz*. This sequence was not present in either of the Neurospora mapping projects (MIPS *N. crassa* database or Fungal Genome Resource at http://gene.genetics.uga.edu/).

DISCUSSION

Many aspects of the developmental cycle of *N. crassa* are under the control of the circadian clock (Loros 1998). As was shown previously, clock control of gene expression can be observed in submerged shaken cultures, although development is suppressed under these conditions (Loros *et al.* 1989). cDNAs were therefore sequenced from two mycelial libraries that correspond to two different times within the circadian day (Bell-Pedersen *et al.* 1996) to identify genes that are expressed under these conditions and to find new clock-controlled genes.

TABLE 3

Summary of newly identified clock-controlled genes

Gene	Contig[a]	Clone used as probe	Peak time	Light induction	Transcript size in kb	Linkage group[b]
A. Time of mRNA peak, transcript size, and chromosomal assignments						
ccg-13	1432	a5f7nm	CT 0-1	No	0.6	I
ccg-14	1421	a341nm	CT 0-1	No	1.2	V
ccg-15	1411	a8d1ne	CT 4-6	No	1.6	V
lyz	1442	a2f10nm	CT 2-4	No	1.0	?

Gene	Homolog, organism (accession no.)	*P/E*
B. Sequence homologs found using BLASTX		
ccg-13	Phase specific protein, *A. dermatitidis* (GenBank accession no. AF277086)	$3e^{-10}$
ccg-14	snodprot1, *P. nodoroum* (GenBank accession no. AF074941)	$7e^{-39}$
ccg-15	SPS2 homologue, *S. cerevisiae* (Saccharomyces Genome Database via RefSeq, NP_009634)	$9e^{-45}$
lyz	Lysozyme, *Chalaropsis* sp. (Swissprot accession no. P00721)	$4e^{-85}$

[a] A single contig identification number is given in those cases in which several contigs have homology to the same gene. This can happen in cases when one contig corresponds to the 5′ end and one to the 3′ end of a given gene. A complete list of contigs corresponding to a single gene can be found at the Genetics website at http://www.genetics.org/supplemental/, at http://www.genome.ou.edu/fungal.html/, and at http://www.dartmouth.edu/~jdunlap/.

[b] Assignment to linkage groups was done by comparison to databases of different mapping projects; for more information see text.

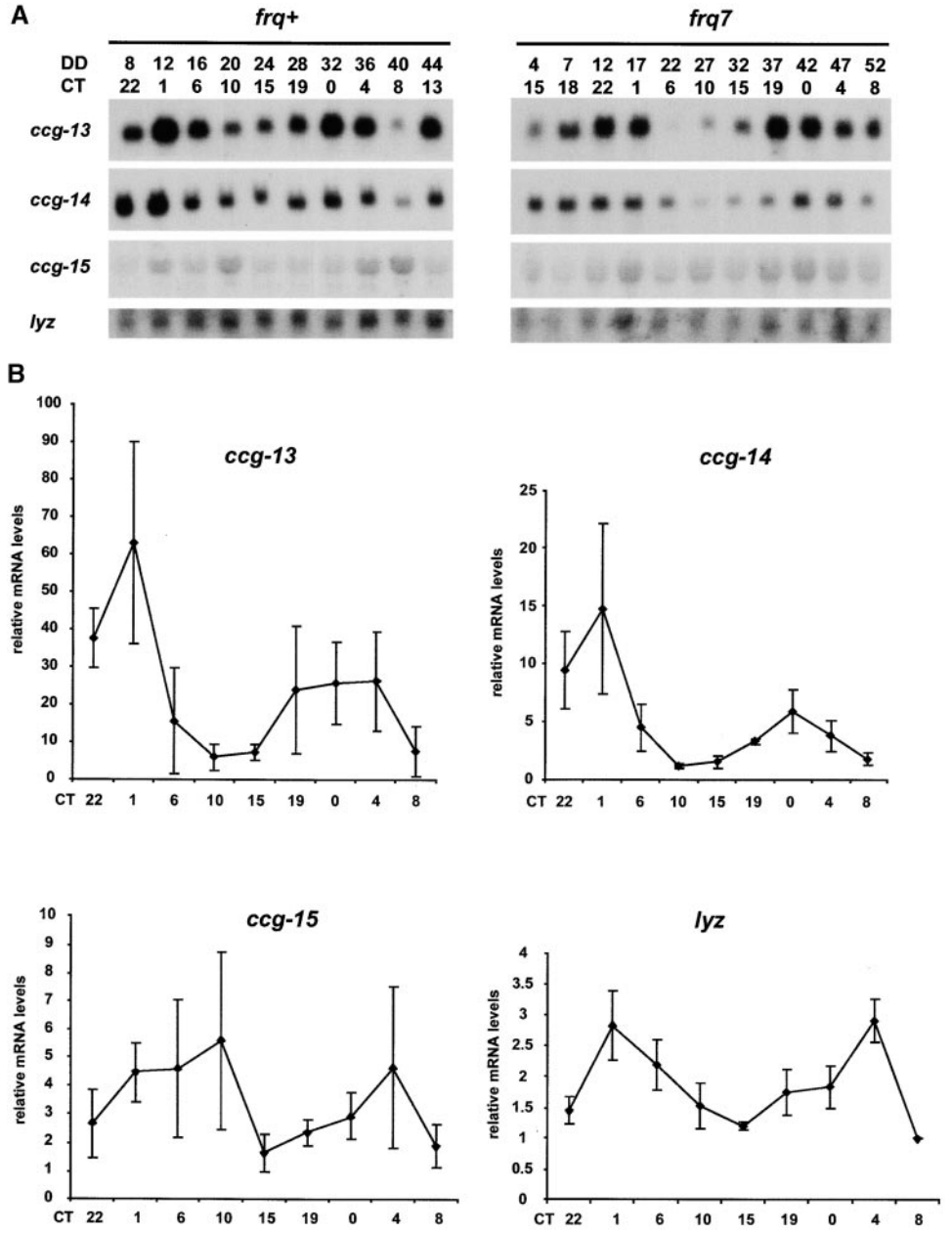

FIGURE 2.—Rhythmic transcript accumulation of newly identified *ccgs* in *frq*$^{+}$ and *frq*^{7} (29-hr period) strains. (A) Northern blot hybridization of total RNA (40 μg) extracted from *frq*$^{+}$ and *frq*^{7} strains after growth in darkness for the indicated constant darkness (DD) times. The corresponding CT (circadian time) is given below the DD values. Probes used for hybridization are indicated on the left. (B) Densitometry of *ccg* mRNAs. Normalization was to ethidium-bromide-stained rRNAs on the gels. The lowest value was set to 1. For each graph, at least three independent experiments were used for the calculations of averages and error bars, two of which were conducted with a *frq*$^{+}$ strain, and one with a *frq*^{7} strain. (In the case of *lyz*, the last three data points consist of only two independent experiments, one with a *frq*$^{+}$ strain and one with a *frq*^{7} strain.)

A large number of previously unknown genes were identified by sequencing cDNAs from two mycelial libraries of *N. crassa*: The RNAs used as templates for the construction of the cDNA libraries sequenced in these studies were derived from cultures held under starvation conditions in the dark for 43 hr (BELL-PEDERSEN *et al.* 1996). cDNA libraries sequenced in previous studies in Neurospora correspond to genes expressed in vegetative mycelial, conidiating, or sexually developing cultures (NELSON *et al.* 1997; DOLAN *et al.* 2000). Thus, because the growth conditions were distinct, we expected to identify a large number of novel genes, and this appears to be the case. Among the 1431 genes that were identified within the 20,000 sequences generated, 721 (50%) did not show any homology to previously identified genes within the NCBI database, whereas 710 genes have matches among the sequences within the NCBI nonredundant databases, as of November 2000. As the total number of Neurospora genes is estimated to be between 10,000 and 14,000 (KELKAR *et al.* 2001), it is surprising that the number of expressed genes identified here represents only 10% of the potentially expressed sequences of *N. crassa.* However, the ESTs generated within this project provide a valuable tool for fungal researchers and can be used for further studies of gene expression, including as probes for microarrays in transcriptional profiling experiments.

The relative abundance of cDNAs in the two libraries does not necessarily reflect gene expression patterns *in vivo*: To our knowledge, there have been few reports of

"digital Northerns" that were confirmed by conventional methods at a larger scale (see, however, Prade *et al.* 2001). Hwang and co-workers identified genes involved in cardiac hypertrophy (Hwang *et al.* 2000), and Huang *et al.* (1999) found several genes differentially expressed in prostate cancer by EST profiling. Both groups confirmed their results using Dot Blot and reverse transcription PCR techniques to survey the expression of 22 and 7 candidate genes, respectively. In an approach to construct the transcriptional profile for skeletal muscle, Bortoluzzi *et al.* (2000) found that electronic expression profiles correlated with serial analysis of gene expression results for overall expression levels of the highly expressed genes in muscle. Efforts have been made to investigate the genes for plant lipid biosynthesis and to compare electronic expression profiles with enzyme activities *in vivo* (Mekhedov *et al.* 2000). The authors find similarities as well as differences between *in silico* and *in vivo*, but as enzyme activities do not necessarily reflect mRNA expression patterns, the reasons for these differences remain to be clarified.

In this investigation, hybridization of RNA time courses with 26 probes derived from cDNA clones representing different contigs did not reveal a strong correlation between the relative abundance of a cDNA among the EST sequences and the expression pattern of the gene *in vivo*. Of the 19 genes with a high probability for differential expression (Audic and Claverie 1997), three turned out to be under clock control and showed peak mRNA levels in the subjective morning, whereas the other 16 did not reveal consistent patterns of differential expression (Table 1). Among the two genes with probabilities for differential expression between 0.95 and 0.999, none is under clock control (Table 1). Screening of 5 genes that displayed probabilities of <0.95 for differential expression did reveal one clock-controlled gene (Table 1), which might indicate that these significance levels lead to many false-positive but few false-negative results. Overall, the number of false positives is relatively high, being $\sim$82%, which does not indicate a strong correlation between EST abundance and *in vivo* expression.

There are several possible explanations for this finding. One might be that the number of sequences obtained is still too low to result in a representative picture of gene expression. The number of genes encoded by the Neurospora genome has been estimated between 10,000 and 14,000 (Nelson *et al.* 1997; Kelkar *et al.* 2001); therefore, sequencing 13,000 cDNA clones might not be enough to make reliable predictions about the situation *in vivo*. This possibility is strengthened by the finding that the five out of seven previously identified *ccgs*, which are predicted correctly by the digital Northern (Table 2), are among the *N. crassa* genes with the highest expression levels *in vivo* as well as the highest count numbers in the libraries. This might also indicate that our EST data are valuable to estimate overall expression levels, similar to investigations done by other groups (Bortoluzzi *et al.* 2000; Mekhedov *et al.* 2000), but that the more subtle changes regulated by the circadian clock are not yet visible among the number of ESTs generated to date. Also to be taken into account is the fact that genes that are neither under clock control nor under any stringent control mechanism under the chosen growth conditions may display random fluctuations of expression. This could mimic the peak and trough patterns of clock-controlled genes when viewed only at two time points.

Another issue to be considered is the evaluation of the statistical significance of the data available. Most of the genes used as probes displayed differences of counts in the morning *vs.* the evening libraries that indicate differential expression with probability levels of >0.999 according to formulas developed to evaluate the outcome of electronic transcriptional profiling (Audic and Claverie 1997). However, this significance level might still be too high given that Hwang *et al.* (1999) used significance levels <0.0002 to identify putative candidate genes for differential expression in cardiac hypertrophy and still found the ratio of false positives to be 10–30%. Also, the formula does not require a minimal number of EST counts, which might be required to make valuable predictions (see above). Plus, the ratios were calculated with the assumption that the overall number of mRNAs present in the mycelium is the same in the morning and the evening. This assumption may not be correct, as Martens and Sargent (1974) found that overall RNA contents in Neurospora mycelium grown on solid medium cycle rhythmically over the circadian day, although with low amplitude. The authors found that the amplitude of the RNA rhythm was even lower in liquid medium, but as they did not use starvation conditions, this might not be true for the mycelia that were used for the libraries in the present study. Also, their study dealt with overall RNA content, not with mRNA.

If the number of mRNAs is greatly different in the morning *vs.* the evening, then sequencing about the same number of cDNA clones for both libraries will not give an accurate picture. Also, the presence of multiple cDNAs for highly expressed genes (such as *ccg-1*, *ccg-2*, and *ccg-7*; see Table 2) may lead to an underrepresentation of genes with lower expression, especially at the time when the highly expressed genes reach their peaks; the 20 genes with the highest count numbers constitute 32% of all ESTs in the morning library and only 11% of all ESTs in the evening library (data not shown). Another caveat might be that some of the clones in the libraries were sequenced from both ends and others only from one end, meaning that some clones would be counted twice. But as this occurs randomly within the whole data set, it may be assumed that this would not influence the overall outcome of the analysis. Still another reason might be that the two libraries were

made from different strains (the morning library from a *frq*^{7} strain and the evening library from a *frq*$^+$ strain; Bell-Pedersen *et al.* 1996) to harvest tissues at the same developmental stage but at different times within the circadian cycle. Although the strains were believed to be isogenic, except for the *frq* allele, the formal possibility exists that strain differences might contribute to the divergence of sequencing data from *in vivo* expression, although until now, no difference in overall expression levels between *frq*$^+$ and *frq*^{7} strains was observed (Bell-Pedersen *et al.* 1996).

Four new clock-controlled genes were identified: Hybridization of rhythmic RNA time courses with probes representing different contigs resulted in the identification of four new clock-controlled genes, *ccg-13*, *ccg-14*, *ccg-15*, and *lyz*. The new *ccgs* vary greatly in their amplitudes and robustness of rhythm (Figure 2). Since the signals and the robustness of the rhythm seen with *ccg-15* and *lyz* were significantly reduced, there might be some doubts about the extent of rhythmicity. Nevertheless, these data are reported as they may indicate different levels of clock control. The Neurospora clock is thought to exert its function over a variety of different pathways and is also interconnected with other levels of regulation such as nutrition and development (Loros 1998; Dunlap 1999). Therefore, it is likely that a variety of *ccgs* are expressed with different degrees of amplitude and robustness of rhythm. The new *ccgs* peak in the early subjective morning, as do previously identified clock-controlled genes (Bell-Pedersen *et al.* 1996). *ccg-13* shows similarity to a phase-specific protein from *A. dermatitidis* (GenBank accession no. AF277086), *ccg-14* is strongly homologous to snodprot1 from the ascomycete *P. nodorum* (GenBank accession no. AF074941), and *ccg-15* displays strong sequence similarity to a sporulation-specific protein (SPS2) homolog from *S. cerevisiae* (Saccharomyces Genome Database via RefSeq, accession no. NP_009634), but the actual function of all three genes in Neurospora remains to be elucidated. *lyz* is quite similar to lysozyme from the ascomycete *Chalaropsis* sp. (Swissprot accession no. P00721) and therefore almost certainly encodes the corresponding enzyme in Neurospora; it might be involved in defense mechanisms. The fact that 4 out of 26 genes tested displayed rhythmic expression confirms the prediction that a significant proportion of *N. crassa* genes is under clock control. Future studies may reveal additional clock-influenced gene expression patterns in Neurospora.

The authors thank Dr. Christian Heintzen for sharing materials and stimulating discussions. This research was supported by grants from the National Institute of Health (R37-GM 34985 to J.C.D. and MH44651 to J.C.D. and J.J.L.), the National Science Foundation (MCB-0084509 to J.J.L.), the Norris Cotton Cancer Center core grant at Dartmouth Medical School, and by a grant from the National Science Foundation EPSCoR program to B.A.R. M.N. received an Emmy-Noether-Fellowship from the German Science Foundation (Deutsche Forschungsgemeinschaft, Bonn-Bad Godesberg, Germany).

LITERATURE CITED

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:** 3389–3402.

Audic, S., and J.-M. Claverie, 1997 The significance of digital gene expression profiles. Genome Res. **7:** 986–995.

Bell-Pedersen, D., M. L. Shinohara, J. J. Loros and J. C. Dunlap, 1996 Circadian clock-controlled genes isolated from *Neurospora crassa* are late night to early morning specific. Proc. Natl. Acad. Sci. USA **93:** 13096–13101.

Birnboim, H. C., and J. Doly, 1979 A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucleic Acids Res. **7:** 1513–1522.

Bodenteich, A., S. Chissoe, Y. F. Wang and B. A. Roe, 1994 Shotgun cloning as the strategy of choice to generate templates for high-throughput dideoxynycleotide sequencing, pp. 42–50 in *Automated DNA Sequencing and Analysis Techniques*, edited by M. D. Adams, C. Fields and J. C. Venter. Academic Press, London.

Boguski, M. S., T. M. Lowe and C. M. Tolstoshev, 1993 dbEST—database for "expressed sequence tags." Nat. Genet. **4:** 332–333.

Bortoluzzi, S., F. d'Alessi, C. Romualdi and G. A. Danieli, 2000 The human adult skeletal muscle transcriptional profile reconstructed by a novel computational approach. Genome Res. **10:** 144–349.

Davis, R. L., and D. deSerres, 1970 Genetic and microbial research techniques for *Neurospora crassa*. Methods Enzymol. **27A:** 79–143.

Davis, R. W., 2000 *Neurospora: Contributions of a Model Organism*. Oxford University Press, New York.

Dolan, P. L., D. O. Natvig and M. A. Nelson, 2000 Neurospora proteome 2000. Fungal Genet. Newsl. **47:** 7–24.

Duggan, D., M. Bittner, Y. Chen, P. Meltzer and J. Trent, 1999 Expression profiling using cDNA microarrays. Nat. Genet. **21:** 10–14.

Dunlap, J. C., 1999 Molecular bases for circadian clocks. Cell **96:** 271–290.

Ewing, B., L. Hillier, M. C. Wendl and P. Green, 1998 Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. **8:** 175–185.

Felch, J., T. Inagami and J. Hash, 1975 The *N, O*-diacetylmuramidase of *Chalaropsis species*. V. The complete amino acid sequence. J. Biol. Chem. **250:** 3713–3720.

Hwang, D. M., A. A. Dempsey, C.-Y. Lee and C.-C. Liew, 2000 Identification of differentially expressed genes in cardiac hypertrophy by analysis of expressed sequence tags. Genomics **66:** 1–14.

Huang, G. M., W. Ng, J. Farkas, L. He, H. A. Liang *et al.*, 1999 Prostate cancer expression profiling by cDNA sequencing analysis. Genomics **59:** 178–186.

Kelkar, H. S., J. Griffith, M. E. Case, S. F. Covert, R. D. Hall *et al.*, 2001 The *Neurospora crassa* genome: cosmid libraries sorted by chromosome. Genetics **157:** 979–990.

Loros, J. J., 1998 Time at the end of the millennium: the Neurospora clock. Curr. Opin. Microbiol. **1:** 698–706.

Loros, J. J., S. A. Denome and J. C. Dunlap, 1989 Molecular cloning of genes under the control of the circadian clock in Neurospora. Science **243:** 385–388.

Maniatis, T., E. F. Fritsch and J. Sambrooke, 1982 *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Martens, C. L., and M. L. Sargent, 1974 Circadian rhythms of nucleic acid metabolism in *Neurospora crassa*. J. Bacteriol. **117:** 1210–1215.

Mekhedov, S., O. Martinez de Ilarduya and J. Ohlrogge, 2000 Toward a functional catalog of the plant genome. A survey of genes for lipid biosynthesis. Plant Physiol. **122:** 389–401.

Nelson, M. A., S. Kang, E. Braun, M. Crawford, P. Dolan *et al.*, 1997 Expressed sequences form conidial, mycelial, and sexual stages of Neurospora. Fungal Genet. Biol. **21:** 348–363.

Ohlrogge, J., and C. Benning, 2000 Unraveling plant metabolism by EST analysis. Curr. Opin. Plant Biol. **3:** 224–228.

Pandey, A., and F. Lewitter, 1999 Nucleotide sequence databases: a gold mine for biologists. Trends Biochem. Sci. **24:** 276–280.

Pazzagli, L., G. Cappugi, M. Giampaolo, G. Camici, A. Santini *et al.*, 1999 Purification, characterization and amino acid sequence of

cerato-platanin, a new phytotoxic protein from *Ceratocystis fimbriata* f. sp. *platani*. J. Biol. Chem. **274:** 24959–24964.

Prade, R. A., P. Ayoubi, S. Krishnan, S. Macwana and H. Russell, 2001 Accumulation of stress and inducer-dependent plant-cell-wall-degrading enzymes during asexual development in *Aspergillus nidulans*. Genetics **157:** 957–967.

Rosenblum, B. B., L. G. Lee, S. L. Spurgeon, S. H. Khan, S. M. Menchen *et al.*, 1997 New dye-labeled terminators for improved DNA sequencing patterns. Nucleic Acids Res. **25:** 4500–4504.

White, O., and A. R. Kerlavage, 1996 TDB: new databases for biological discovery. Methods Enzymol. **266:** 27–40.

Yarden, O., M. Plamann, D. Ebbole and C. Yanofsky, 1992 *cot-1*, a gene required for hyphal elongation in *Neurospora crassa* encodes a protein kinase. EMBO J. **11:** 2159–2166.

Communicating editor: J. Arnold